



E-ISSN 3032-601X & P-ISSN 3032-7105

Vol. 3, No. 1, 2026

MISTER

**Journal of Multidisciplinary Inquiry in Science,
Technology and Educational Research**

**Jurnal Penelitian Multidisiplin dalam Ilmu
Pengetahuan, Teknologi dan Pendidikan**

**UNIVERSITAS SERAMBI MEKKAH
KOTA BANDA ACEH**

mister@serambimekkah.ac.id

Journal of Multidisciplinary Inquiry in Science Technology
and Educational Research

Journal of MISTER

Vol. 3, No. 1, 2026

Pages: 1603–1611

Klasifikasi Topik Berita Online Bahasa Indonesia Menggunakan *Support Vector Machine* Berbasis TF-IDF

I Gede Angga Pratama, I Wayan Sudiarsa, Dewa Made Artha Wiguna,
I Made Dwi Windu Saputra, I Made Teguh Winantara

Institut Bisnis dan Teknologi Indonesia, Denpasar

Article in Journal of MISTER

Available at : <https://jurnal.serambimekkah.ac.id/index.php/mister>

DOI : <https://doi.org/10.32672/mister.v3i1.4130>

How to Cite this Article

APA : Pratama, I. G. A., Sudiarsa, I. W., Wiguna, D. M. A., Saputra, I. M. D. W., & Winantara, I. M. T. (2026). Klasifikasi Topik Berita Online Bahasa Indonesia Menggunakan Support Vector Machine Berbasis TF-IDF. Journal of Multidisciplinary Inquiry in Science, Technology and Educational Research, 3(1), 1603 - 1611. <https://doi.org/10.32672/mister.v3i1.4130>

Others Visit : <https://jurnal.serambimekkah.ac.id/index.php/mister>

MISTER: Journal of Multidisciplinary Inquiry in Science, Technology and Educational Research is a scholarly journal dedicated to the exploration and dissemination of innovative ideas, trends and research on the various topics include, but not limited to functional areas of Science, Technology, Education, Humanities, Economy, Art, Health and Medicine, Environment and Sustainability or Law and Ethics.

MISTER: Journal of Multidisciplinary Inquiry in Science, Technology and Educational Research is an open-access journal, and users are permitted to read, download, copy, search, or link to the full text of articles or use them for other lawful purposes. Articles on Journal of MISTER have been previewed and authenticated by the Authors before sending for publication. The Journal, Chief Editor, and the editorial board are not entitled or liable to either justify or responsible for inaccurate and misleading data if any. It is the sole responsibility of the Author concerned.



Klasifikasi Topik Berita Online Bahasa Indonesia Menggunakan Support Vector Machine Berbasis TF-IDF

I Gede Angga Pratama^{1*}, I Wayan Sudiarsa^{2*}, Dewa Made Artha Wiguna³,
I Made Dwi Windu Saputra⁴, I Made Teguh Winantara⁵
Rekayasa Sistem Komputer, Institut Bisnis dan Teknologi Indonesia, Denpasar, Indonesia^{1,2,3,4,5}

Email : 1gedeanggapratama977@gmail.com, 2sudiarsa@instiki.ac.id,
3dewaguna022@gmail.com, 4windusaputra129@gmail.com, 5teguhwinantara05@gmail.com
*Email Korespondensi: sudiarsa@instiki.ac.id

Diterima: 10-01-2026

| Disetujui: 20-01-2026

| Diterbitkan: 22-01-2026

The development of online media has led to a significant increase in the number of online news stories, necessitating an automated method to efficiently group and label news topics. The manual labeling process is time-consuming and resource-intensive and prone to inconsistencies. Therefore, this study aims to implement and evaluate the Support Vector Machine (SVM) algorithm in automating the labeling of Indonesian-language online news topics. This study uses a quantitative approach with an experimental method on an online news dataset that has undergone text cleaning and preprocessing stages, including normalization, tokenization, stopword removal, and advanced text processing. Document feature representation is performed using the Term Frequency-Inverse Document Frequency (TF-IDF) method to convert text data into high-dimensional numeric vectors. A classification model is built using a Support Vector Machine with a linear kernel and implemented using the Scikit-learn library. The dataset is divided into training data and test data with a ratio of 80% and 20% to evaluate model performance objectively. The test results show that the SVM model is able to achieve an accuracy level of 79.38%, with relatively balanced precision, recall, and F1-score values across most topic classes. Confusion matrix analysis shows that most documents were correctly classified, although errors still occurred in topics with similar contexts. The findings of this study indicate that the combination of TF-IDF and Support Vector Machines is effective for classifying Indonesian online news texts and has the potential to support the development of a digital content management system based on topic labeling automation.

Keywords: text classification; support vector machine; TF-IDF; online news; Indonesian language.

ABSTRAK

Perkembangan media daring menyebabkan peningkatan jumlah berita online yang signifikan, sehingga diperlukan metode otomatis untuk membantu pengelompokan dan pelabelan topik berita secara efisien. Proses pelabelan manual membutuhkan waktu dan sumber daya yang besar serta rentan terhadap inkonsistensi. Oleh karena itu, penelitian ini bertujuan untuk menerapkan dan mengevaluasi algoritma Support Vector Machine (SVM) dalam melakukan otomatisasi pelabelan topik berita online berbahasa Indonesia. Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen terhadap dataset berita online yang telah melalui tahap pembersihan dan preprocessing teks, meliputi normalisasi, tokenisasi, penghapusan stopword, dan pengolahan teks lanjutan. Representasi fitur dokumen dilakukan menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF) untuk mengubah data teks menjadi vektor numerik berdimensi tinggi. Model klasifikasi dibangun menggunakan Support Vector

Machine dengan kernel linear dan diimplementasikan menggunakan pustaka Scikit-learn. Dataset dibagi menjadi data latih dan data uji dengan perbandingan 80% dan 20% untuk mengevaluasi kinerja model secara objektif. Hasil pengujian menunjukkan bahwa model SVM mampu mencapai tingkat akurasi sebesar 79,38%, dengan nilai precision, recall, dan F1-score yang relatif seimbang pada sebagian besar kelas topik. Analisis confusion matrix menunjukkan bahwa sebagian besar dokumen berhasil diklasifikasikan dengan benar, meskipun masih terdapat kesalahan pada topik-topik yang memiliki kemiripan konteks. Temuan penelitian ini menunjukkan bahwa kombinasi TF-IDF dan Support Vector Machine efektif digunakan untuk klasifikasi teks berita online Bahasa Indonesia dan berpotensi mendukung pengembangan sistem manajemen konten digital berbasis otomatisasi pelabelan topik.

Katakunci: klasifikasi teks; support vector machine; TF-IDF; berita online; bahasa Indonesia.

PENDAHULUAN

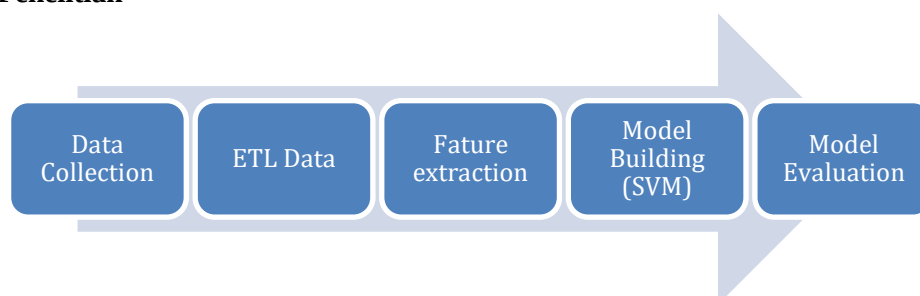
Pesatnya perkembangan teknologi informasi dan internet telah mendorong pertumbuhan media berita daring secara signifikan. Media berita online mempublikasikan ribuan artikel setiap hari dengan beragam topik, seperti politik, ekonomi, sosial, dan olahraga. Kondisi ini menyebabkan proses pengelompokan dan pelabelan topik secara manual menjadi tidak efisien serta berpotensi menimbulkan inkonsistensi dalam pengelolaan konten digital. Ketidakkonsistenan pelabelan topik dapat berdampak pada menurunnya kualitas sistem pencarian, pengarsipan, dan penyajian informasi, sehingga diperlukan pendekatan otomatis yang lebih andal dan terukur (Saigal, 2020; Yang & Liu, 2016).

Sejumlah penelitian dalam sepuluh tahun terakhir menunjukkan bahwa pendekatan machine learning efektif digunakan dalam klasifikasi teks berita. Penelitian oleh Kadhim (2019) menunjukkan bahwa representasi fitur berbasis Term Frequency–Inverse Document Frequency (TF-IDF) mampu merepresentasikan karakteristik dokumen teks secara efektif. Sementara itu, Suyanto et al. (2018) serta Hidayat dan Wibowo (2021) melaporkan bahwa algoritma Support Vector Machine (SVM) memiliki kinerja yang stabil dan unggul dalam menangani data teks berdimensi tinggi, termasuk pada klasifikasi berita daring. Temuan serupa juga ditunjukkan pada penelitian klasifikasi berita dan judul berita online menggunakan kombinasi TF-IDF dan SVM (Putra & Fauzi, 2023; Zulfikar & Nugroho, 2022). Namun demikian, sebagian besar penelitian tersebut masih berfokus pada dataset berbahasa asing atau belum secara spesifik membahas otomatisasi pelabelan topik berita online berbahasa Indonesia.

Dalam konteks tersebut, text mining menjadi pendekatan yang relevan untuk mengekstraksi informasi dari data teks tidak terstruktur, khususnya melalui tugas klasifikasi teks. Proses klasifikasi teks umumnya melibatkan tahapan preprocessing teks, ekstraksi fitur menggunakan TF-IDF, serta penerapan algoritma klasifikasi seperti SVM (Arifin et al., 2021; Kadhim, 2019). Berdasarkan tinjauan literatur tersebut, terdapat celah penelitian dalam penerapan algoritma Support Vector Machine untuk otomatisasi pelabelan topik berita online berbahasa Indonesia yang mempertimbangkan karakteristik bahasa dan konteks lokal. Oleh karena itu, penelitian ini bertujuan untuk menerapkan dan mengevaluasi kinerja algoritma Support Vector Machine dengan representasi fitur TF-IDF dalam mengotomatisasi pelabelan topik berita online, guna meningkatkan efisiensi dan konsistensi pengelolaan konten digital.

METODE PENELITIAN

1. Desain Penelitian



Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen pada data teks berita online berbahasa Indonesia. Penelitian bertujuan untuk mengembangkan dan mengevaluasi model klasifikasi otomatis pelabelan topik berita menggunakan algoritma Support Vector Machine (SVM).

Tahapan penelitian meliputi pengumpulan data, preprocessing teks, ekstraksi fitur menggunakan Term Frequency–Inverse Document Frequency (TF-IDF), pemodelan klasifikasi, serta evaluasi kinerja model menggunakan metrik akurasi, precision, recall, dan F1-score.

2. Data dan Sumber Data

Data yang digunakan dalam penelitian ini merupakan dataset berita online berbahasa Indonesia yang bersifat dataset sekunder dan dapat diakses secara publik. Dataset diperoleh dari final_merge_dataset/kaggle yang berisi 80480 dokumen berita dengan atribut judul, isi berita, dan label topik. Dataset dapat diakses melalui <https://www.kaggle.com/datasets/sh1zuka/indonesia-news-dataset-2024>. Data yang tidak memiliki label topik atau memiliki label kosong dieliminasi pada tahap awal untuk memastikan kualitas data.

3. Preprocessing dan Ekstraksi Fitur

Tahap preprocessing dilakukan untuk meningkatkan kualitas data teks, meliputi penghapusan data duplikat, normalisasi teks, tokenisasi, penghapusan stopword Bahasa Indonesia, dan stemming. Selanjutnya, data teks direpresentasikan dalam bentuk numerik menggunakan metode TF-IDF untuk menghasilkan vektor fitur berdimensi tinggi.

4. Pemodelan Klasifikasi

Model klasifikasi yang digunakan adalah Support Vector Machine (SVM) dengan kernel linear. Dataset dibagi menjadi data latih dan data uji dengan perbandingan 80%:20%. Implementasi model dilakukan menggunakan library Scikit-learn dengan kelas LinearSVC dan parameter regularisasi default.

5. Evaluasi Model

Evaluasi kinerja model dilakukan menggunakan metrik akurasi, precision, recall, dan F1-score. Selain itu, confusion matrix digunakan untuk menganalisis distribusi kesalahan klasifikasi pada setiap kelas topik.

6. Perangkat dan Lingkungan Eksperimen

Eksperimen dilakukan menggunakan bahasa pemrograman Python pada lingkungan Google Colaboratory dengan bantuan pustaka Pandas dan Scikit-learn.

HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil eksperimen otomatisasi pelabelan topik berita online Bahasa Indonesia menggunakan algoritma Support Vector Machine (SVM). Pembahasan meliputi proses pengumpulan data, hasil preprocessing, kinerja model klasifikasi, serta interpretasi hasil berdasarkan konsep dasar dan penelitian sebelumnya.

1. Proses Pengumpulan dan Karakteristik Data

1.1. Pengumpulan Dataset Berita Online

Dataset yang digunakan dalam penelitian ini merupakan kumpulan berita online berbahasa Indonesia yang diperoleh dari beberapa media daring. Proses pengumpulan data dilakukan secara tidak langsung dengan memanfaatkan dataset sekunder yang telah tersedia dalam format digital.

Setelah proses seleksi dan pembersihan data, diperoleh sebanyak 6.161 dokumen berita yang memiliki label topik yang valid dan siap digunakan dalam proses pemodelan. Perbedaan penulisan label topik disebabkan oleh variasi penamaan pada dataset asli.

Tabel 1. Sepuluh Topik Berita Dengan Jumlah Data Terbanyak

index	Topik Berita	Jumlah Data
0	Jakarta	136
1	prabowo subianto	132
2	kpk	112
3	prabowo	103
4	Prabowo	96
5	Pramono Anung	72
6	KPK	64
7	info-tempo	62
8	Ridwan Kamil	56
9	Jokowi	54

(Sumber: Data diolah penulis (2026))

1.2. Pembagian Data Latih dan Data Uji

Dataset yang telah melalui tahap preprocessing selanjutnya dibagi menjadi data latih dan data uji dengan perbandingan 80% data latih dan 20% data uji. Pembagian ini bertujuan untuk memastikan bahwa model dapat dievaluasi secara objektif pada data yang belum pernah dilihat sebelumnya.

2. Hasil Preprocessing dan Ekstraksi Fitur

2.1. Hasil Preprocessing Teks

Tahapan preprocessing dilakukan untuk meningkatkan kualitas data teks sebelum proses klasifikasi. Proses ini meliputi normalisasi teks, tokenisasi, penghapusan stopwords, dan stemming Bahasa Indonesia. Hasil preprocessing menunjukkan bahwa teks berita menjadi lebih ringkas dan representatif, sehingga dapat mengurangi noise dan meningkatkan efektivitas proses ekstraksi fitur.

2.2. Representasi Fitur Menggunakan TF-IDF

Setelah preprocessing, dokumen teks direpresentasikan dalam bentuk vektor numerik menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Metode ini menghasilkan matriks fitur berdimensi tinggi yang mencerminkan tingkat kepentingan setiap kata dalam dokumen.

Penggunaan TF-IDF terbukti efektif dalam merepresentasikan karakteristik dokumen berita dan mendukung kinerja algoritma SVM dalam proses klasifikasi.

3. Hasil Klasifikasi Menggunakan Support Vector Machine

3.1. Kinerja Model Klasifikasi

Model Support Vector Machine (SVM) dilatih menggunakan data latih dan diuji menggunakan data uji. Evaluasi kinerja model dilakukan menggunakan metrik akurasi, precision, recall, dan F1-score.

Hasil evaluasi menunjukkan bahwa model SVM mampu mencapai akurasi sebesar 0,7938, dengan nilai precision, recall, dan F1-score yang relatif seimbang pada masing-masing kelas topik.

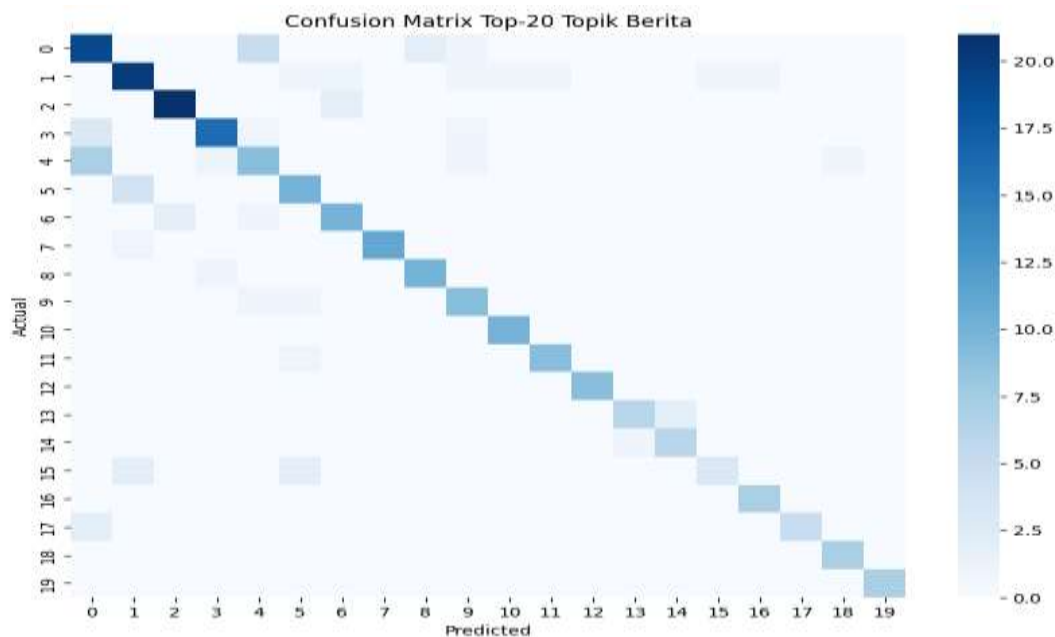
Tabel 2. nilai precision, recall, dan F1-score

	precision	recall	f1-score	support
Jakarta	0.740741	0.740741	0.740741	27
KPK	0.769231	0.769231	0.769231	13
Prabowo	0.529412	0.473684	0.5	19
Pramono Anung	0.666667	0.714286	0.689655	14
Ridwan Kamil	0.692308	0.818182	0.75	11
banjir	0.909091	1	0.952381	10
gaza	0.75	0.857143	0.8	7
gempa	1	1	1	9
info-tempo	1	0.916667	0.956522	12
israel	0.857143	0.75	0.8	8
jokowi	0.833333	0.909091	0.869565	11
judi online	1	0.714286	0.833333	7
kebakaran	0.9	0.9	0.9	10
kecelakaan	0.875	1	0.933333	7
kemensos	0.875	1	0.933333	7
kpk	0.913043	0.913043	0.913043	23
mpr	1	1	1	7
prabowo	0.888889	0.761905	0.820513	21
prabowo subianto	0.612903	0.703704	0.655172	27
pramono anung	0.75	0.428571	0.545455	7
accuracy	0.793774	0.793774	0.793774	0.793774
macro avg	0.828138	0.818527	0.818114	257
weighted avg	0.798005	0.793774	0.792261	257

Sumber: Data diolah penulis (2026)

3.2. Analisis Confusion Matrix

Confusion matrix digunakan untuk menganalisis kesalahan klasifikasi yang terjadi pada masing-masing kelas topik. Berdasarkan hasil confusion matrix, sebagian besar dokumen berhasil diklasifikasikan dengan benar, meskipun masih terdapat beberapa kesalahan klasifikasi pada topik yang memiliki kemiripan konteks.



Gambar 1. Confusion Matrix Hasil Klasifikasi Top-20 Topik Berita Online Bahasa Indonesia Menggunakan Support Vector Machine Berbasis TF-IDF
(Sumber: Data diolah penulis (2026))

Gambar 1 memperlihatkan confusion matrix hasil klasifikasi top-20 topik berita online menggunakan algoritma Support Vector Machine berbasis TF-IDF. Sumbu vertikal menunjukkan label aktual, sedangkan sumbu horizontal menunjukkan label hasil prediksi. Dominasi nilai pada diagonal utama menandakan bahwa sebagian besar data berhasil diklasifikasikan dengan benar. Kesalahan klasifikasi relatif kecil dan umumnya terjadi pada topik yang memiliki kemiripan konteks, sehingga menunjukkan bahwa model SVM memiliki performa yang baik dalam melakukan otomatisasi pelabelan topik berita.

4. Pembahasan dan Keterkaitan dengan Penelitian Terdahulu

Hasil penelitian ini menunjukkan bahwa algoritma Support Vector Machine memiliki performa yang baik dalam melakukan pelabelan topik berita online Bahasa Indonesia. Temuan ini sejalan dengan penelitian sebelumnya yang menyatakan bahwa SVM efektif dalam menangani data teks berdimensi tinggi dan bersifat sparse.

Penggunaan TF-IDF sebagai metode ekstraksi fitur juga berkontribusi positif terhadap kinerja model, sebagaimana dilaporkan dalam berbagai studi klasifikasi teks sebelumnya. Dengan demikian, hasil penelitian ini memperkuat temuan terdahulu dan menunjukkan bahwa pendekatan SVM masih relevan dan kompetitif dalam konteks klasifikasi berita online.

5. Implikasi Hasil Penelitian

Secara teoritis, penelitian ini memperkuat konsep penggunaan machine learning dalam text mining, khususnya klasifikasi teks berita. Secara praktis, hasil penelitian ini dapat dimanfaatkan oleh pengelola media daring sebagai dasar pengembangan sistem otomatisasi pelabelan topik berita guna meningkatkan efisiensi pengelolaan konten digital.

KESIMPULAN

Penelitian ini berhasil menerapkan algoritma Support Vector Machine berbasis pembobotan TF-IDF untuk melakukan otomatisasi pelabelan topik berita online berbahasa Indonesia. Hasil pengujian menunjukkan bahwa model mampu mengklasifikasikan topik berita dengan performa yang baik, yang ditunjukkan oleh dominasi prediksi benar pada confusion matrix serta nilai akurasi yang tinggi pada data uji. Temuan ini membuktikan bahwa kombinasi TF-IDF dan SVM efektif digunakan untuk klasifikasi teks berdimensi tinggi pada domain berita online.

Meskipun demikian, penelitian ini memiliki keterbatasan, terutama pada distribusi data yang tidak merata antar topik dan adanya kemiripan konteks antar kelas yang menyebabkan kesalahan klasifikasi pada beberapa topik tertentu. Oleh karena itu, penelitian selanjutnya disarankan untuk menggunakan teknik penyeimbangan data, mengeksplorasi pendekatan multi-label classification, serta membandingkan kinerja SVM dengan metode lain seperti Random Forest atau model berbasis deep learning guna memperoleh hasil klasifikasi yang lebih optimal.

DAFTAR PUSTAKA

- Arifin, N., Enri, U., & Sulistiyowati, N. (2021). Penerapan support vector machine dengan TF-IDF n-gram untuk text classification. *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, 6(2), 120–129. <https://doi.org/10.30998/string.v6i2.10125>
- Dadgar, S. M. H., Araghi, M. S., & Farahani, M. M. (2016). A novel text mining approach based on TF-IDF and support vector machine for news classification. *Proceedings of the 2016 IEEE International Conference on Engineering and Technology (ICETECH)* (pp. 112–117). IEEE. <https://doi.org/10.1109/ICETECH.2016.7569223>
- Hidayat, S., & Wibowo, A. T. (2021). Analisis kinerja support vector machine pada klasifikasi teks berdimensi tinggi. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 5(1), 78–85. <https://doi.org/10.29207/resti.v5i1.2865>
- Husain, N. P., Sukirman, S., & Sajiah, S. A. J. (2024). Analisis sentimen ulasan pengguna TikTok menggunakan TF-IDF dan support vector machine. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 11(1), 45–54. <https://doi.org/10.25126/jtiik.20241101>
- Ionendri, N. A., Candra, F., & Rizal, A. (2025). Online news classification using TF-IDF and natural language processing. *Jurnal Aplikasi dan Teori Ilmu Komputer (JACOST)*, 4(1), 1–10.
- Kadhim, A. I. (2019). Term weighting for feature extraction on Twitter: A comparison between TF-IDF and TF-RF. *International Journal of Advanced Computer Science and Applications*, 10(2), 1–7. <https://doi.org/10.14569/IJACSA.2019.0100201>
- Lestari, V. B. (2025). Evaluation of TF-IDF feature extraction techniques in sentiment analysis. *Jurnal Komputer dan Aplikasi*, 13(2), 87–95.
- Putra, P. R. B., & Fauzi, M. A. (2023). Klasifikasi judul berita online menggunakan metode support vector machine dan TF-IDF. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 7(3), 1345–1352.
- Saigal, P. (2020). Multi-category news classification using support vector machine. *SN Applied Sciences*, 2(3), 1–9. <https://doi.org/10.1007/s42452-020-2170-5>
- Suyanto, S., Prasetyo, E., & Nugroho, A. (2018). Text classification using support vector machine with TF-

- IDF weighting. *International Journal of Artificial Intelligence Research*, 2(1), 20–27.
- Suroyo, H., & Pramono, A. (2025). Comparison of text representation methods for sentiment classification. *Journal of Artificial Intelligence and Information Technology*, 9(1), 55–63.
- Wang, S., & Manning, C. D. (2017). Baselines and bigrams: Simple, good sentiment and topic classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (pp. 90–94). <https://doi.org/10.18653/v1/P12-2018>
- Yang, Y., & Liu, X. (2016). A re-examination of text categorization methods. *ACM SIGIR Forum*, 50(2), 3–14. <https://doi.org/10.1145/2970398.2970401>
- Zhang, X., Zhao, J., & LeCun, Y. (2017). Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 30, 649–657.
- Zulfikar, M., & Nugroho, L. E. (2022). Klasifikasi berita hoaks berbahasa Indonesia menggunakan TF-IDF dan support vector machine. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, 11(4), 387–394.